Salience-based reinforcement of a spiking neural network leads to increased syllable production

Anne S. Warlaumont
Cognitive and Information Sciences, University of California, Merced, Merced, California 95343
Email: awarlaumont2@ucmerced.edu, Web: annewarlaumont.org

***Abstract*—Canonical babbling is vocal babbling that contains syllabic patterning like that in adult speech. Its emergence during the first year of human infancy is one of the most significant pre-speech vocal motor milestones. This paper focuses on a spiking neural network model that controls the lip and jaw muscles of an articulatory speech synthesizer and learns to produce canonical babbling. The model was adapted to receive reinforcement when it produced a sound with high auditory salience. Salience-reinforced versions of the model increased their rates of canonical babbling over the course of learning more than their yoked controls. This supports the idea that both intrinsic reinforcement and social reinforcement both contribute to human acquisition of canonical babbling.**

I. Introduction

A. The emergence of syllable production in infancy

The emergence of canonical babbling is one of the most significant early milestones in infants' speech development. Canonical babbling is defined as the production of canonical syllables, either alone or in sequence. A canonical syllable is a syllable containing both a consonant and a full vowel where the transition between the consonant and vowel is rather fast, which is characteristic of syllables in adult speech [1]. In typically developing children this ability appears between 4 and 10 months of age [2]. Canonical babbling can take the form of a single canonical syllable, of repetitions of the same consonant-vowel permutation (a.k.a. reduplicated babbling), or of a sequence of different consonant-vowel permutations (a.k.a. variegated babbling).

How do infants come to learn to produce canonical babbling? Much previous research has focused on the role that social reinforcement might play. If we assume that a caregiver knows that some sounds are higher quality (e.g., containing not just vowel sounds but also adult-like consonants) than others, then they can choose to selectively reinforce infants' productions of those sounds. Infants can then use that selective (i.e. contingent) reinforcement as a cue to guide their learning with the result being that they will come to produce more of the reinforced sound. We know from observational studies of human children that parents do respond differently to infant vocalizations that contain consonants than to those that do not contain consonants [3] and that contingent responding to infant speech-related vocalizations leads to increases in infants' rates of production of speech-related vocalizations [4].

It has also been proposed that an intrinsic interest in the vocalizations one produces could play a role in shaping pre-speech vocal development [1], [5]–[8]. One specific idea is that infant vocalizations that are particularly salient to the auditory or somatosensory perceptual systems could be in and of themselves rewarding to infants [5], [8]. Another possibility is that infants might be motivated by internal goals to produce sounds that expand their repertoires [6]. The fact that infants often vocalize on their own, not necessarily in interaction with any other person [1], supports this role of intrinsic reinforcement in infant vocal learning. Figure 1 illustrates schematically what is meant by social reinforcement and intrinsic reinforcement.

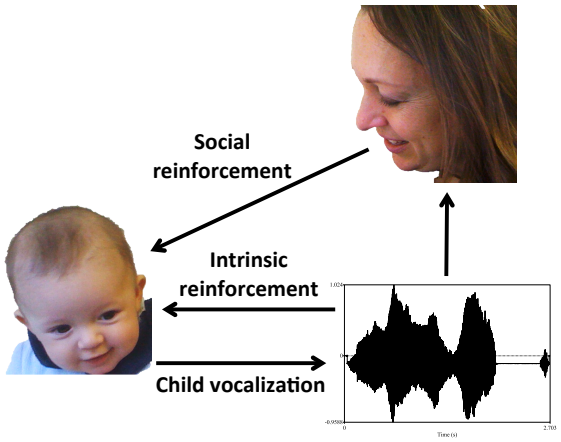

Fig. 1. Illustration of how social reinforcement and intrinsic reinforcement could be involved in infant vocal learning.

The goal of the current study is to determine if auditory salience as the sole reinforcer, i.e. the sole source of feedback, would be sufficient in a neural network model of canonical babbling development. Another goal is to explicitly compare intrinsic, salience-based reinforcement to social reinforcement and see if one is more effective than the other.

B. Previous computational models

There are already a number of computational models of infant vocal development. Most have focused on the development of specific vowel, and occasionally consonant, sounds, focusing on the role of mapping between one's own motor actions and the auditory and somatosensory consequences of those sounds and on the role played by input from another individual, such as a parent. For example, a number of models develop sensory-motor associations in order to learn to imitate vowel sounds [9]–[17]. Some have addressed how we might learn what phonemic categories correspond to what motor commands [12], [14]. And some have addressed how individuals' repertoires of vowel and consonant sounds could be shaped through a combination of exploration and

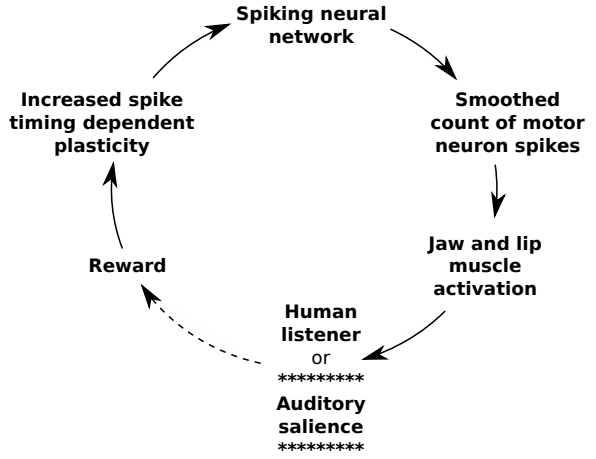

Fig. 2. Overview of the model used here and in [7]. The difference is that here auditory salience forms the basis for reinforcement whereas in the earlier model, a human listener decided when to reinforce the model.

selective responding by other individuals [5], [8]. The next few paragraphs will focus on reviewing in more detail a selection of other models of particular relevance to the present paper, including one model addressing the development of canonical babbling and two addressing the role of intrinsic reinforcement in the development of vowel and consonant inventories. Some work in other motor domains that demonstrates the potential power of intrinsic reinforcement will also be reviewed.

A previous model described in [7] is unique in that it attempts to explain how the canonical babbling milestone might be achieved. It uses a spiking neural network to control a simulated vocal tract, utilizing social reinforcement from a human listener (See Fig. 2). A spiking neural network was chosen because it intrinsically contains complex temporal dynamics [18], [19] without having to a priori set a syllabic rhythmic frame for articulators, as has previously been done in many models that learn to produce both consonants and vowels [5], [12]. After every 5 s of neural simulation, jaw and lip muscle activations were generated using the previous 1 s of neural spiking activity of a subset of neurons designated as motor neurons. These jaw and lip muscle activations were fed into an articulatory synthesizer [20], which generated a sound. A human listened to each sound the model generated and decided whether or not to reinforce it; the listener's goal was to get the model to produce more advanced canonical babbling more often. When the model was reinforced, this led to an increase in the dopamine concentration in the spiking neural network, which temporarily increased the rate of neural network learning. Over the course of two hours of social interaction with the model, the model's performance gradually improved compared to a control model. This result suggests a specific neural mechanism by which reinforcement can lead to the development of canonical babbling and fits with previous research with human infants showing that social reinforcement from a caregiver can lead to vocal learning. A natural question is whether intrinsic, non-social, reinforcement through the same neural reinforcement learning mechanism could also lead to the emergence of canonical babbling.

One vocal development model that is noteworthy for including intrinsic reinforcement is Howard and Messum's model [5]. A combination of both sensory, including auditory and somatosensory, salience and social reinforcement was used by the model in acquiring a repertoire of consonant and vowel sounds. Salience was based on a combination of total acoustic power, whether or not a touch (e.g., the lips touching each other) was perceived, and spectral balance in the acoustic signal. Salience was then combined with a measure of the diversity a sound added to the repertoire and a measure of the effort required to produce the sound. This combination formed an objective function that the model used to explore its vocalization abilities and to add sounds to its permanent vocal repertoire. After this stage of intrinsically guided exploration and repertoire expansion, social reinforcement from a human listener was used to refine the model's repertoire, shaping it to include primarily sounds that are present in English. Note that Howard and Messum's model was not a neural network model, in contrast to the model presented below, and thus addressed mechanisms of learning at a different level of analysis. Also note that the focus was on modeling the acquisition of specific consonant and vowel sounds assuming a set timing for when consonants and vowels should occur and not on the emergence of syllabic patterning itself.

Work by Moulin-Frier and Oudeyer [6] also utilized intrinsically guided learning for vocal motor development, specifically the development of an inventory of vowels. That model explored a space of seven upper vocal tract motor articulators and observed the auditory consequences of each motor exploration. It did not explore in a random way but instead chose a sensory goal and attempted to produce motor commands that would reach that goal. The most sophisticated version of the model chose its sensory goal at each time point to be one which had not already been mastered but for which significant learning progress was possible. Compared to a model that explored randomly, the intrinsically motivated (i.e. "goal selection with reaching") model ended up with a broader, more complete vowel inventory. This work thus demonstrates that intrinsically guided learning can be quite powerful. Note though that this approach has also not yet been used to address the emergence of canonical babbling, i.e. babbling with adult-like syllabic patterning, and also does not operate at a neural level of analysis. Also, it focuses on intrinsic adaptive goal selection, which is a different aspect of intrinsically guided learning than what will be focused on in the model presented below, which focuses on using salience as an intrinsic reward.

Although there has been only a little work thus far on intrinsically guided learning in vocal motor development, there has been some additional work in other domains of action control. For instance, a similar approach to that taken by Moulin-Frier and Oudeyer has been applied in a more general motor learning context in which it was possible for a robot to move its limbs and mouth and to produced a fixed set of sounds. The robot was placed on a real human infant play mat and was given an "adult" robot to interact with. The robot demonstrated a developmental sequence that nicely corresponds to what is displayed by human infants [21].

Additionally, of particular interest for the present paper, Lewis et al. [22] showed that visual salience as the sole reinforcer in a reinforcement learning model of eye movement was sufficient for the model to develop the ability to jointly attend to objects with a simulated caregiver. This study shows

the power of salience as an intrinsic reward and was a direct inspiration for the work presented below.

II. Method

The neural network model was the same spiking neural network as that used in [7] and was an adaptation of that in [23]. The spiking neural network contained 1000 spiking neurons including 200 inhibitory and 800 excitatory. 50 of the excitatory neurons were randomly chosen to be motor neurons. At the start of each simulation, each neuron was randomly connected to 100 other neurons. Inhibitory neurons could not be connected to other inhibitory neurons but all other connection types were allowed.

The simulations ran in 1 ms time steps. At each time step, each neuron received current from each neuron that connected to it that had spiked during the previous millisecond, with that current being proportional to the connection weight from the presynaptic neuron to the potsynaptic neuron. At each time step each neuron also received some random quantity of current input. Current inputs increased the voltage of the neuron. Counteracting these inputs was a tendency for each neuron's voltage to decay over time. Neuron spikes were defined as occurring whenever a a voltage above 30 mV; after each spike a neuron's voltage was reset to -65 mV.

Every 5 s of simulated neural activity, the preceding 1 s of motor neurons' spikes were summed then smoothed by performing a 100 ms moving average. The resulting 900 ms time series was multiplied by 2.5 and fed to an articulatory synthesizer as masseter (a muscle that closes the jaw) and orbicularis oris muscle activations [20]. Contraction of the masseter muscle closes the jaw and contraction of the orbicularis oris muscle closes the lips. Lung volume and laryngeal muscle activations were preset to ensure phonation; only the mouth and jaw muscles changed across different vocalizations. Because masseter and orbicularis oris muscles were the only muscles manipulated all consonants produced by the model were similar phonetically speaking and tended to sound something like [w]. After each sound was produced the model was either reinforced or not reinforced.

Reinforcement, when received, increased the dopamine concentration in the neural network, increasing the rate of spike timing dependent plasticity (STDP), a type of learning used with spiking neural networks that mimics that which has been observed at real cortical synapses. Each synapse's eligibility to be strengthened was always stored. The eligibility trace was increased when the postsynaptic neuron fired soon after the presynaptic neuron, with this increase being larger the sooner after the presynaptic neuron that the postsynaptic neuron fired. Every 10 ms, the current dopamine concentration was multiplied by each synapse's eligibility trace and synapse strengths were increased by the resulting amounts. Both the dopamine concentration and the synaptic eligibility traces decayed exponentially over time. Thus, reward-modulated STDP biased the model toward producing neuronal activations similar to those produced prior to the reward, thereby biasing the network to produce vocalizations more like the one it had just produced. Only synapses with excitatory presynaptic neurons were allowed to change—the strength of inhibitory-to-excitatory synapses was held constant throughout each simulation.

The difference between this model and that in [7] was the source of reinforcement. Instead of having a human listen to each sound the model made and decide whether or not to reinforce it, reinforcement was given if the estimated auditory salience of the model's vocalization was above a threshold (see Fig. 2). If the estimated auditory salience of a vocalization was below the threshold, reinforcement was withheld.

Auditory salience was estimated using Coath et al.'s biologically inspired Auditory Salience Model [24], [25]. The method first takes a sound waveform and processes it using 30 filters modeling the processing that occurs at the cochlea. The resulting cochlear representation of the sound is then processed in various additional ways modeling the transformations of the representation of the sound in the central nervous system, including auditory cortex. Activation of the filters that represent cortical sound processing can be tracked over time, and rises and falls can be detected. Greater changes over time, that is greater rises and falls, of the activation of these cortical filters are assumed to correspond to points of higher auditory salience. The algorithm has previously been successfully applied to modeling tapping along to music [24].

Two examples of muscle activation timeseries and corresponding vocalization waveforms, cochlear responses, and auditory salience time series for two vocalizations produced by the human-reinforced version of the babbling model [7] are given in Figure 3. The salience trace for the first 150 ms of each 900 ms vocalization was ignored since it represented salience associated simply with the onset of sound and the wish here was to focus on auditory features that were potentially related to the syllabic patterning of a sound. We took the time series, $s(t)$, computed using Coath et al.'s auditory saliency algorithm and representing the estimated salience of a vocalization, v, took the absolute value of each element in that salience time series, and then summed all the values in the time series to get a single salience value for the whole vocalization, $S(v)$:

$$S(v) = \sum_{t=151\text{ms}}^{900\text{ms}} |s(v,t)| \tag{1}$$

Thus, $S(v)$ operationalizes auditory salience as the magnitude of temporal variance in cortical filter activations.

We ran two versions of the model, a low threshold version in which the salience value of a vocalization had to be greater than 4.5 in order for the sound to be reinforced and a high threshold version in which the salience value had to be greater than 7.5. These two values were chosen after listening to a number of synthesized vocalizations and comparing their $S(v)$ values to their perceived canonical babble quality. Nine simulations of each version were run. In addition, eighteen yoked control simulations were run, one for each of the eighteen salience-based reinforcement simulations. Each yoked control was run using a new random initialization of neural connections and new random inputs to each neuron at each time step. It therefore received reinforcement that was not contingent on anything having to do with the acoustics of its vocalization but rather it received reinforcement at the same exact times as one of the real, salience based reinforcement simulations.

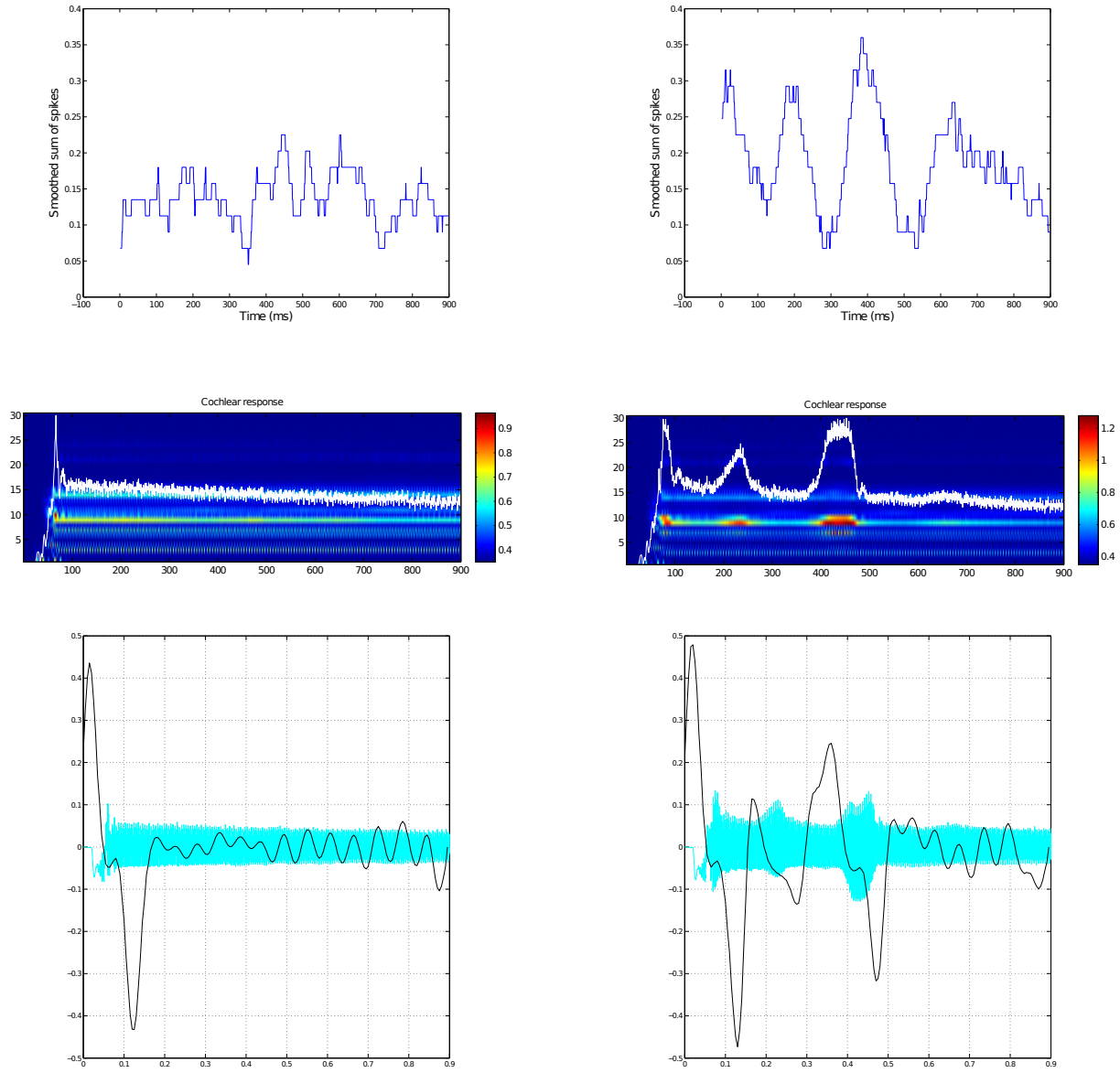

Fig. 3. Two examples of model vocalizations. On the left is a non-canonical vocalization that did not contain any syllables, only an extended vowel sound. On the right is a vocalization that contained two well-formed syllables and that would be classified as canonical babble. Top: the lip and jaw muscle activations that generated the sounds. Middle: simulated cochlear responses to the sounds. Bottom: In cyan is the vocalization waveform and in black is the salience trace for the sound. The salience is based on detected changes in activation of modeled auditory cortex filters. The salience is 10.83 for the canonical sound shown on the right, which is larger than the salience of the non-canonical sound, which is 3.50.

Each simulation was run for 2 hours of simulated time, so that 1,440 $(2 \cdot 60 \cdot 60/5)$ vocalizations were produced during the course of each simulation. Each vocalization sound made by every model was saved as a .wav file and its salience value, $S(v)$, was also saved so that we could measure the simulations' performance over the course of training. Of interest in addition to tracking the increase or lack thereof in salience over learning, was how the salience-reinforced simulations' vocalizations would sound to human listeners compared to the previous human-reinforced version of the model. Thus, two human listeners, each of whom had received some training listening to human infant examples of canonical babbling and of non-canonical vocalizations, were asked to listen to vocalizations produced by the first two low threshold simulations, the first two high threshold simulations, and the human reinforced simulation from [7]. Time constraint was the reason only a subset of the simulations were included. The sounds produced across learning for the five simulations were combined into a single list and then the order was scrambled so the listeners did not know which simulation had produced a sound nor when during learning it had been produced. The listeners rated each sound on a four-point scale where a score of 1 was given to sounds with no well-formed syllables, i.e. non-canonical vocalizations, and a score of 4 was given to

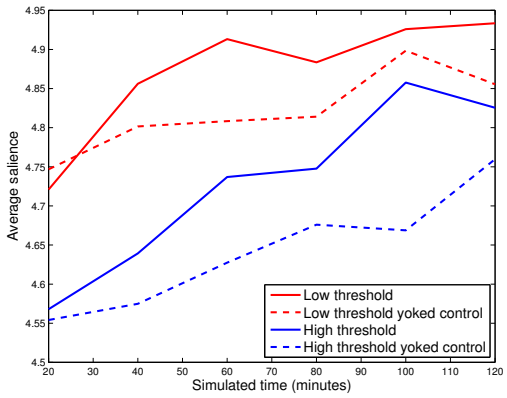

Fig. 4. Change in salience of vocalization produced by the model over time. Lines represent 20-minute averages over nine simulations.

TABLE I. AVERAGE (AND STANDARD ERROR) OF THE AVERAGE SALIENCE, $S(v)$, PRODUCED BY EACH MODEL DURING THE FIRST AND LAST 20 MINUTES OF LEARNING.

	First 20 minutes	Last 20 minutes
Low threshold	4.72 (0.03)	4.93 (0.05)
Low threshold yoked control	4.75 (0.03)	4.85 (0.05)
High threshold	4.57 (0.02)	4.83 (0.05)
High threshold yoked control	4.55 (0.02)	4.76 (0.05)

sounds with multiple very well-formed syllables, i.e. strong examples of speech-like canonical babbling. Each listener rated 7,200 $(5 \cdot 1,440)$ sounds.

III. RESULTS

Figure 4 shows the average salience, $S(v)$, of vocalizations produced over the course of learning for each simulation type. Table I gives the average salience performance for each version of the model both in the first 20 minutes and in the last 20 minutes of learning. Both the salience-reinforced versions and the yoked control versions of the model showed increases in average salience over time, with the salience-reinforced showing greater increases. Over all trials, salience was .05 higher for the low-threshold version simulations than their yoked controls, $p < .001$. Similarly, over all trials, salience was .09 higher for the high-threshold version simulations than their yoked controls, $p < .001$.

The low threshold versions and their yoked controls showed greater increases than the high threshold versions. Over all trials, the low threshold version simulations' salience was .14 higher than the high threshold version simulations' salience. Thus, the low threshold versions, at least over the 2 hour simulation time period, improved more than the high threshold versions, $p = .015$.

Figure 5 shows the average human judgments of the vocalizations produced by two of the low threshold salience reinforced simulations, two of the high threshold salience reinforced simulations, and the human reinforced simulation from [7] over time. All three simulation types' human judgment scores increased over training: for the human reinforced simulation, $\beta = .05$, $p = .015$, for the low threshold version, $\beta = .029$, $p = .031$, and for the high threshold version,

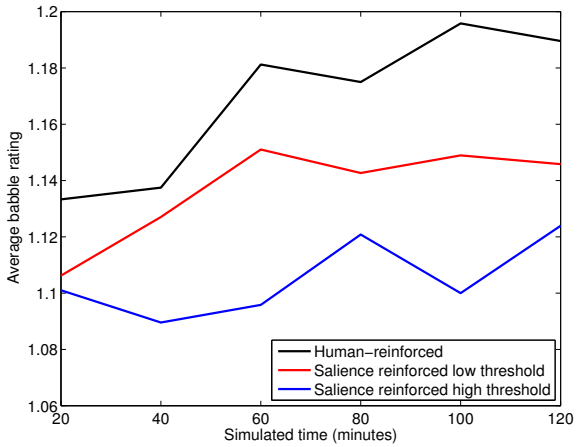

Fig. 5. The human listeners' jugdements regarding the quality of babbling. Lines represent 20-minute averages over the two listeners' judgments for each model type.

$\beta = .026$, $p = .028$. None of the versions performed statistically significantly better than another, though this may be due to lack of statistical power. Qualitatively, the human ratings of the high and low salience versions' vocalizations mirrored the salience values of two versions' vocalizations, with the low threshold version being rated higher than the high threshold version.

IV. DISCUSSION

The results indicate that salience based reinforcement can be a high quality source of feedback in learning to produce canonical babble. Here a spiking neural network that controls the lip and jaw muscles of a vocal tract simulation was selectively reinforced when the vocalizations it produced were deemed auditory salient. Salience was estimated using an algorithm that mimics auditory processing occurring in the cochlea and central nervous system, including auditory cortex. The salience detection algorithm is particularly attuned to temporal changes in auditory stimuli, so that syllabic productions, which have temporally varying properties over time, are perceived as more salient than pure vowel sounds which do not change over time (even if the vowel sounds actually contain more acoustic energy).

The fact that the model increased in quality of vocalization over time suggests that intrinsic, salience-guided learning may be an important complement to socially-guided learning in early vocal development. This is consistent with previous studies of human vocalization which have shown that while social reinforcement does appear to play a role in early vocal learning, there are also reasons to believe that infants sometimes vocalize for its own sake, and not just to elicit a social response. The idea that both intrinsic reinforcement and social reinforcement may play a role in early vocal development also supports an idea proposed by Oller [1] that canonical babbling is such an important behavior that its development has become canalized by evolution. That is, it has been made robust to disturbances. Having multiple pathways, such as an salience-based-reinforcement pathway and a social-reinforcement pathway to the emergence of canonical babbling would be a mechanism for canalization of the development

of this important skill. The reason that severe to profound deafness produces severe delays in canonical babbling whereas many other risk factors do not may be because it affects both social and intrinsic reinforcement, as an infant with little hearing ability will not only lack the ability to hear other individuals' auditory responses (a source of social reinforcement) but also will lack the ability to hear their own vocalizations (a source of intrinsic reinforcement).

Interestingly, in the simulations presented here, there appeared to be an advantage when a low threshold for reinforcement was used vs. a high threshold. The reason for this advantage may be that each simulation was allowed to run only for two hours of simulated time, during which the high threshold version received less frequent reinforcement than the high threshold version. Perhaps when allowed to run for longer, the high salience version would catch up to the low salience version or even show an advantage.

In contrast to the fixed-threshold salience-reinforced simulations, the human listener was presumably adaptive to the model's performance over time when deciding whether to provide reinforce. It seems likely that adaptive learner goals and adaptive reinforcement are the norm in human infancy and will typically lead to better performance by developmental robots and computational models. For example, human mothers' responses to infant behaviors do appear to change depending on infant vocal repertoire size [26]. Additionally, good results have been achieved by computational models that adapt their goals based on their previous performances [5], [6], [21] and computational work has demonstrated that different rewards are optimal in different agent environments [27]. The lack of adaptiveness in the salience-reinforced versions may account for the higher performance of the human-reinforced simulation. Further exploration of various adaptive reward schemes is an important future direction for the present model.

There are some other future directions that would also be interesting to pursue. For instance, it has been known for some time that although severe to profoundly deaf infants don't exhibit canonical babbling until later than typically developing infants, they do eventually babble canonically [1]. If the present model were to be modified to include somatosensory salience in addition to auditory salience, it might be possible to develop a realistic model of canonical babbling in deaf infants. Additionally, because the current version of the model only controls the lips and jaw and even those were always given the same input, it does not yet address variegated babbling nor can it be used to model the order in which different syllable types emerge.

Another very interesting future direction would be to incorporate a broader range of upper and lower vocal tract muscles and see if certain types of consonants and vowels emerge sooner than others. This order could then be compared to the order of emergence of those syllables in human infants' repertoires. If such a model were developed it would also permit us to investigate whether infants' specific ambient language environments might influence which of their own vocalizations. It could be the case that intrinsic reward as well as social reward are both biased toward sounds that are included in the language(s) that the infant is learning. The specific consonants produced in the present model would likely be reinforcing to infants and their caregivers in most language environments, since consonants involving lip and jaw closure, such as [m] and [w], are very common throughout the world's languages. However, infants learning languages that lack either of these consonants might find some of the sounds produced by the model here to be less salient and the sounds might also tend to be rated as lower quality by adult speakers of those languages (all the listeners here were English speakers).

Finally, this work makes the prediction that infants should find canonical babble more acoustically salient than non-syllabic vocalizations. It would be very informative to test this prediction, perhaps using a looking time procedure, which has previously been used to demonstrate infant's preferences for speech over non-speech stimuli [28]. If infants' looking time, a measure of preference, is greater for sounds with higher auditory salience as defined here, that would support the present model; if not, it would provide evidence against it.

ACKNOWLEDGMENT

Thanks is due to Sophia Hyatt, Priscilla Montez, Megan Finnegan, and Jessica Alcantara their help listening to the synthesized vocalizations.

REFERENCES

[1] D. K. Oller, *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[2] D. K. Oller, R. E. Eilers, A. R. Neal, and A. B. Cobo-Lewis, "Late onset canonical babbling: a possible early marker of abnormal development," *American journal of mental retardation: AJMR*, vol. 103, no. 3, pp. 249–263, 1998.

[3] J. Gros-Louis, M. J. West, M. H. Goldstein, and A. P. King, "Mothers provide differential feedback to infants' prelinguistic sounds," *International Journal of Behavioral Development*, vol. 30, no. 6, pp. 509–516, 2006.

[4] M. H. Goldstein, A. P. King, and M. J. West, "Social interaction shapes babbling: testing parallels between birdsong and speech," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 13, pp. 8030–8035, 2003.

[5] I. S. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, no. 1, pp. 85–117, 2011.

[6] C. Moulin-Frier and P.-Y. Oudeyer, "Curiosity-driven phonetic learning," in *Proceedings of the 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 2012.

[7] A. S. Warlaumont, "A spiking neural network model of canonical babbling development," in *Proceedings of the 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 2012.

[8] A. S. Warlaumont, G. Westermann, E. H. Buder, and D. K. Oller, "Prespeech motor learning in a neural network using reinforcement," *Neural Networks*, vol. 38, pp. 64–75, 2013.

[9] Y. Yoshikawa, M. Asada, K. Hosoda, and J. Koga, "A constructivist approach to infants vowel acquisition through mother-infant interaction," *Connection Science*, vol. 15, no. 4, pp. 245–258, 2003.

[10] G. Westermann and E. R. Miranda, "A new model of sensorimotor coupling in the development of speech," *Brain and Language*, vol. 89, no. 2, pp. 393–400, 2004.

[11] P.-Y. Oudeyer, *Self-organization in the evolution of speech*. Oxford, UK: Oxford University Press, 2006, no. 6.

[12] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and Language*, vol. 96, no. 3, pp. 280–301, 2006.

[13] H. Kanda, T. Ogata, T. Takahashi, K. Komatani, and H. Okuno, "Continuous vocal imitation with self-organized vowel spaces in recurrent neural network," in *2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 4438–4443.

[14] B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, vol. 51, pp. 793–809, 2009.

[15] I. Heintz, M. Beckman, E. Fosler-Lussier, and L. Ménard, "Evaluating parameters for mapping adult vowels to imitative babbling," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2009.

[16] A. S. Warlaumont, G. Westermann, and D. K. Oller, "Self-production facilitates and adult input interferes in a neural network model of infant vowel imitation," in *AISB 2011 Computational Models of Cognitive Development*, D. Kazakov and G. Tsoulas, Eds. York, UK: Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 2011, pp. 8–12.

[17] K. Miura, Y. Yoshikawa, and M. Asada, "Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver," *Advanced Robotics*, vol. 26, no. 1-2, pp. 23–44, 2012.

[18] C. T. Kello, J. Rodny, A. S. Warlaumont, and D. C. Noelle, "Plasticity, learning, and complexity in spiking networks," *Critical Reviews in Biomedical Engineering*, vol. 40, no. 6, pp. 501–518, 2012.

[19] C. T. Kello, "Critical branching neural networks." *Psychological Review*, vol. 120, no. 1, pp. 230–254, 2013.

[20] P. Boersma, *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics, 1998. [Online]. Available: http://www.fon.hum.uva.nl/paul/papers/funphon.pdf

[21] F. Kaplan and P.-Y. Oudeyer, "In search of the neural circuits of intrinsic motivation," *Frontiers in Neuroscience*, vol. 1, no. 1, pp. 225–236, 2007.

[22] J. M. Lewis, G. O. Deák, H. Jasso, and J. Triesch, "Building a model of infant social interaction," in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, S. Ohlsson and R. Catrambone, Eds. Austin, TX: Cognitive Science Society, 2010, pp. 278–283. [Online]. Available: http://palm.mindmodeling.org/cogsci2010/papers/0046/index.html

[23] E. M. Izhikevich, "Solving the distal reward problem through linkage of STDP and dopamine signaling," *Cerebral Cortex*, vol. 17, no. 10, pp. 2443–2452, 2007.

[24] M. Coath, S. L. Denham, L. Smith, H. Honing, A. Hazan, P. Holonwicz, and H. Purwins, "An auditory model for the detection of perceptual onsets and beat tracking in singing," in *Neural Information Processing Systems, Workshop on Music Processing in the Brain*, 2007.

[25] S. L. Denham, "Auditory salience model," 2008. [Online]. Available: http://emcap.iua.upf.edu/downloads/content_final/auditory_saliency_model.html

[26] M. H. Goldstein and M. J. West, "Consistent responses of human mothers to prelinguistic infants: the effect of prelinguistic repertoire size." *Journal of comparative psychology*, vol. 113, no. 1, pp. 52–58, 1999.

[27] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg, "Intrinsically motivated reinforcement learning: An evolutionary perspective," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 2, pp. 70–82, 2010.

[28] A. Vouloumanos and J. F. Werker, "Tuned to the signal: the privileged status of speech for young infants," *Developmental Science*, vol. 7, no. 3, pp. 270–276, 2004.